

Federated Space-Time Query for Earth Science Data Using OpenSearch Conventions

ESIP Federated Search Cluster

Chris Lynnes, NASA/GSFC

Bruce Beaumont, University of Alabama - Huntsville

Ruth Duerr, National Snow and Ice Data Center

Hook Hua, NASA/JPL

et al.

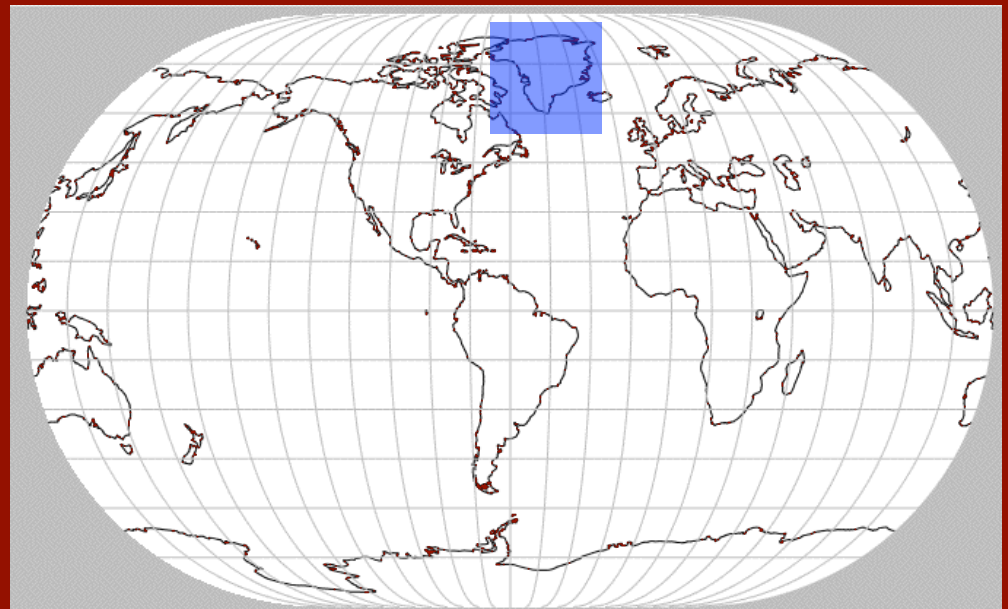
Outline

- Finding Earth science data: why so difficult???
- Space-Time Query with OpenSearch
- Client and server developments

Finding Earth science data: why so difficult???

Many phenomena require space-time searches for distributed data

- E.g., Effect of Arctic Oscillation on precipitation in Greenland
 - GC-Net station data
 - AO indices
 - AIRS atmospheric profiles
 - ECMWF model output
 - NCEP model output, etc.
- Potential data providers:
 - Large data centers
 - Universities
 - Data collection sites
 - Value-added providers
 - Individual investigators



Obtaining satellite data today is tedious, hit-or-miss

Step 1: Search through multiple directories for the right datasets

- “Did I find them all?”

Steps 2-N:

- Foreach data_provider

 - Learn_search_interface()

 - Search_for_data_files()

 - Fetch_data_files()

 - Load_data_into_analysis_tool()

- End foreach

Ideally, you would want your analysis tool to find and fetch data based on the current work context

Space-Time Data Query with OpenSearch

OpenSearch is a simple, extensible, embeddable, machine-callable convention

- www.opensearch.org
 - “a collection of simple formats for the sharing of search results”
- OpenSearch Description Document (XML)
 - Describes a search engine so that it can be used by search clients (incl. Firefox and IE)
 - Specifies syntax for URL-based queries
 - Extensions proposed for Geospatial and Time queries

OpenSearch templates provide the keys to querying heterogeneous search engines

- OpenSearch Description Document includes URL template:

```
<os:Url type="application/atom+xml"      template="http://  
  mirador.gsfc.nasa.gov/cgi-bin/mirador/  
  granlist.pl?dataSet=AIRS2RET.005&page=1 &  
  maxgranules={count}&  
  pointLocation={geo:box}&  
  endTime={time:end}&startTime={time:start}&  
  format=atom">
```

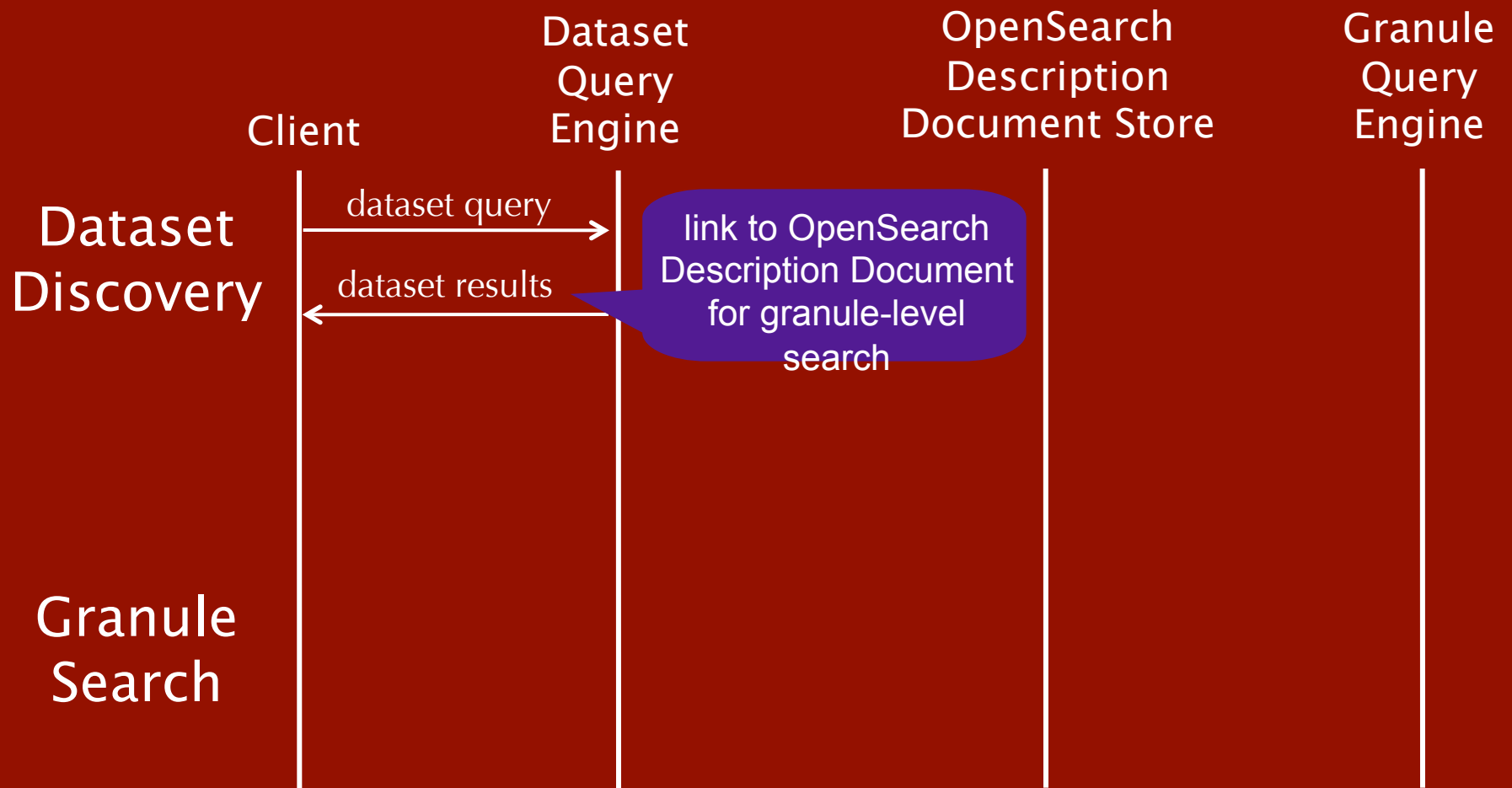
- Just replace placeholders with search criteria and fetch the URL

Data query with space and time works better as a 2-step process

- Search for datasets then granules (files) within selected datasets
- Most dataset-level queries have
 - small results set (dozens)
 - low precision: $\text{precision} = \text{desiderata} / \text{total}$
- Space-time granule queries for a given dataset have
 - large results set (tens of thousands)
 - high precision
- Combining both in one step would produce
 - enormous results set (dozens * tens of thousands)
 - with low precision

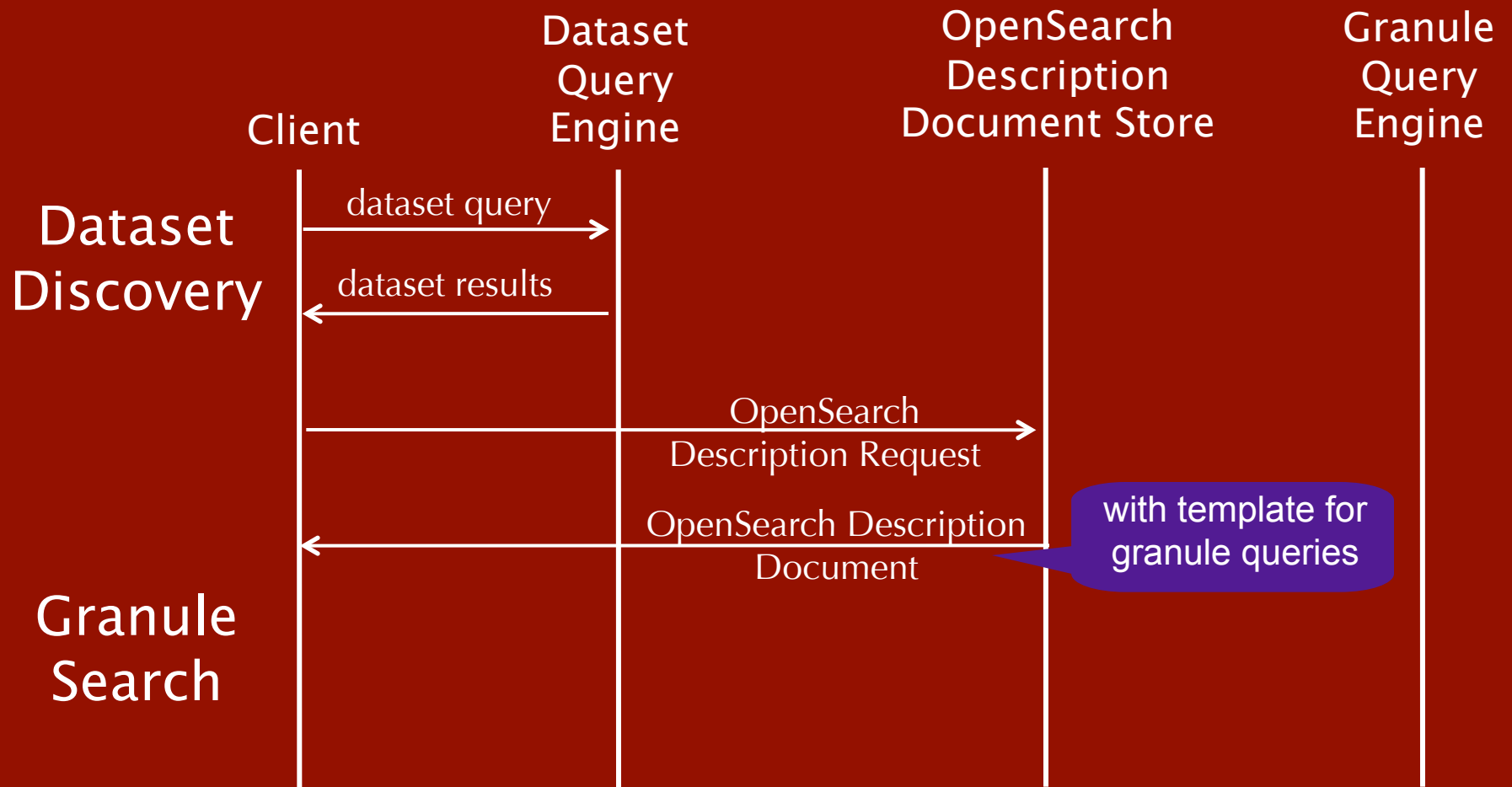
OpenSearch Description Documents provide a path to a recursive two-step search

Recursive OpenSearch begins with a dataset discovery phase

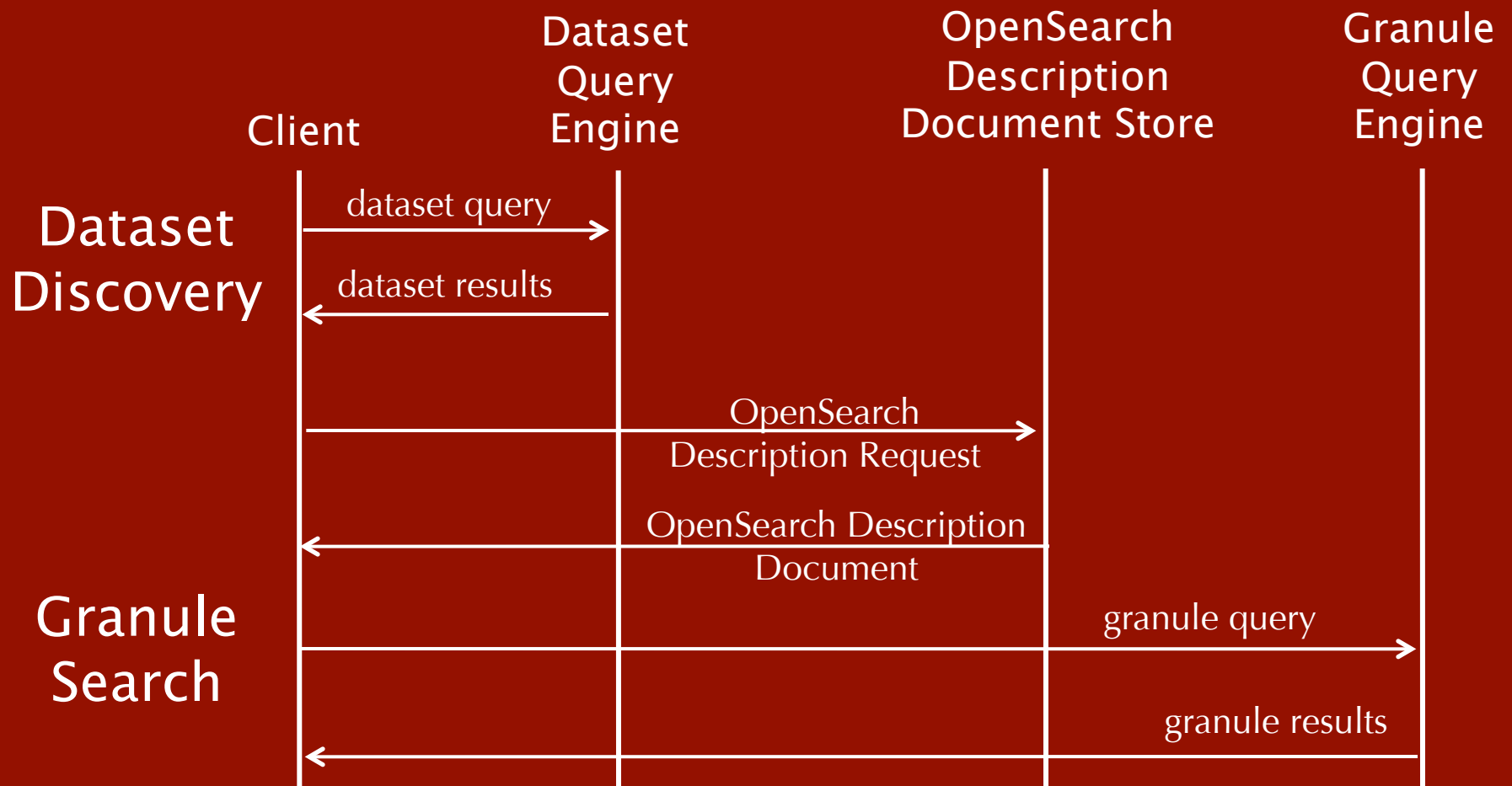


Granule Search

Dataset results link to OpenSearch Description documents



Templates from OpenSearch Description Documents enable granule query construction



The ESIP Federated Search Cluster is defining conventions for a 2-step space time query

- Earth Science Information Partners
 - Consortium of >90 organizations working with remotely sensed Earth observation information
 - Clusters: focus groups to work specific topics
- Federated Search cluster for ESIP community conventions
 - 2-Step (Recursive) OpenSearch

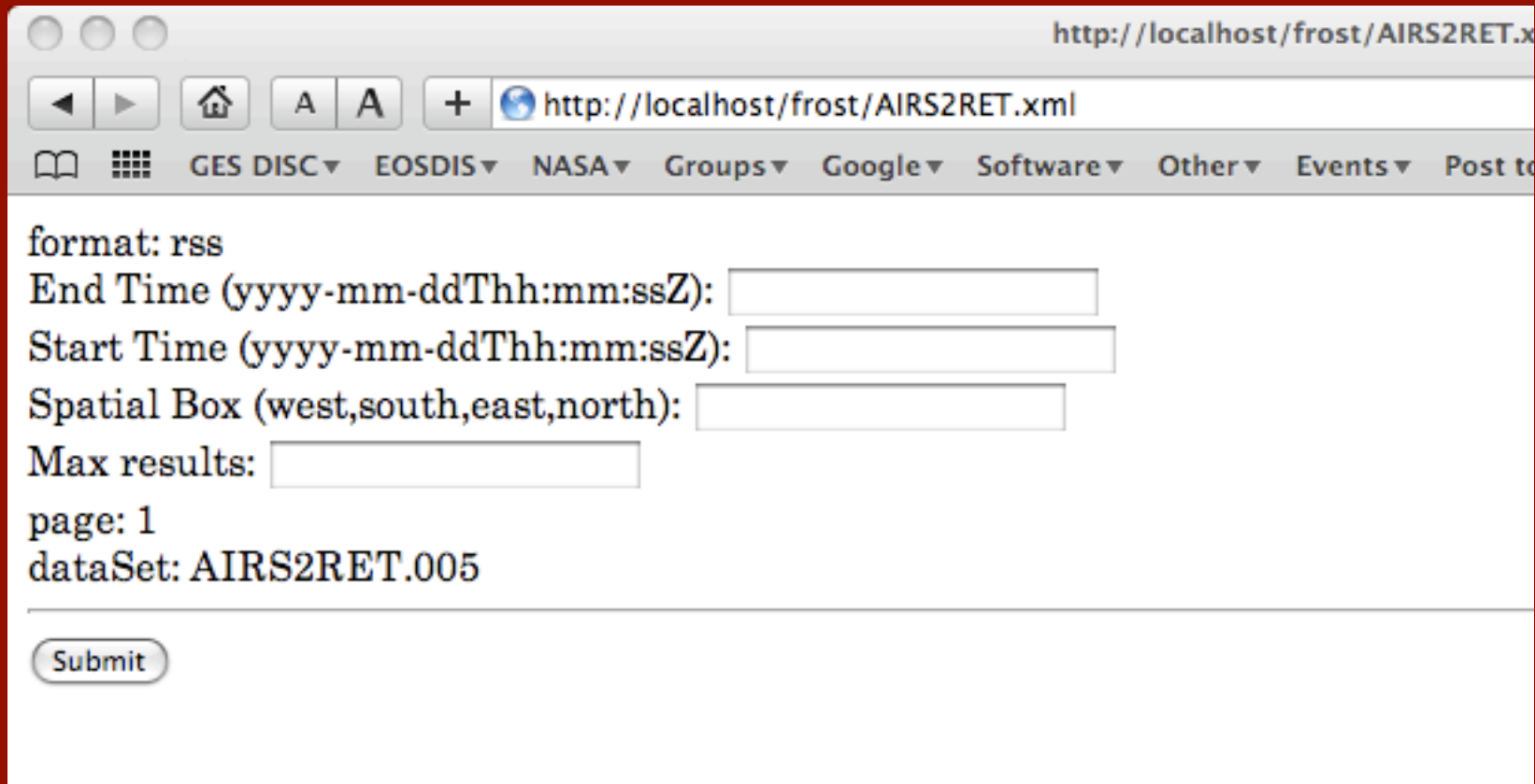
Client and Server Developments

Federated OpenSearch aspects make adoption easier

- Simple / lightweight
- Standards-based, but extensible
- Embeddable
 - In web pages, documents, workflows, analysis tools...

A client can be as simple as an XSLT

- Attach a stylesheet to the OpenSearch Description Document
 - Renders the document in the browser as a search form

A screenshot of a web browser window displaying a search form. The browser's address bar shows the URL "http://localhost/frost/AIRS2RET.xml". Below the address bar is a navigation bar with links: "GES DISC", "EOSDIS", "NASA", "Groups", "Google", "Software", "Other", "Events", and "Post to". The main content area of the browser displays a search form with the following fields and labels: "format: rss", "End Time (yyyy-mm-ddThh:mm:ssZ):", "Start Time (yyyy-mm-ddThh:mm:ssZ):", "Spatial Box (west,south,east,north):", and "Max results:". Each of these labels is followed by a text input field. Below these fields, the text "page: 1" and "dataSet: AIRS2RET.005" is displayed. At the bottom of the form is a "Submit" button.

format: rss

End Time (yyyy-mm-ddThh:mm:ssZ):

Start Time (yyyy-mm-ddThh:mm:ssZ):

Spatial Box (west,south,east,north):

Max results:

page: 1

dataSet: AIRS2RET.005

Several groups are developing servers and clients

- Servers following ESIP Federated Search conventions
 - ACCESS-NEWS
 - EOS Clearinghouse (ECHO)
 - Global Hydrology Resource Center
 - Goddard Earth Sciences Data and Information Services Center (GES DISC)*
 - MODIS Adaptive Processing System
 - National Snow and Ice Data Center
- Clients
 - Mirador (GES DISC)
 - Talkoot (University of Alabama--Huntsville)
 - Reference implementation / test script (GES DISC)*

Future Plans

- Develop / recruit clients
- Support access to Web Services
 - Format conversion, subsetting, OPeNDAP, OGC
 - Servicecasting
 - Atom-based approach to advertising services for ESIP data
- Shrink-wrapped toolset for deploying Recursive OpenSearch servers?

Conclusion

Federated space-time query can be

- lightweight
- inexpensive
- grassroots